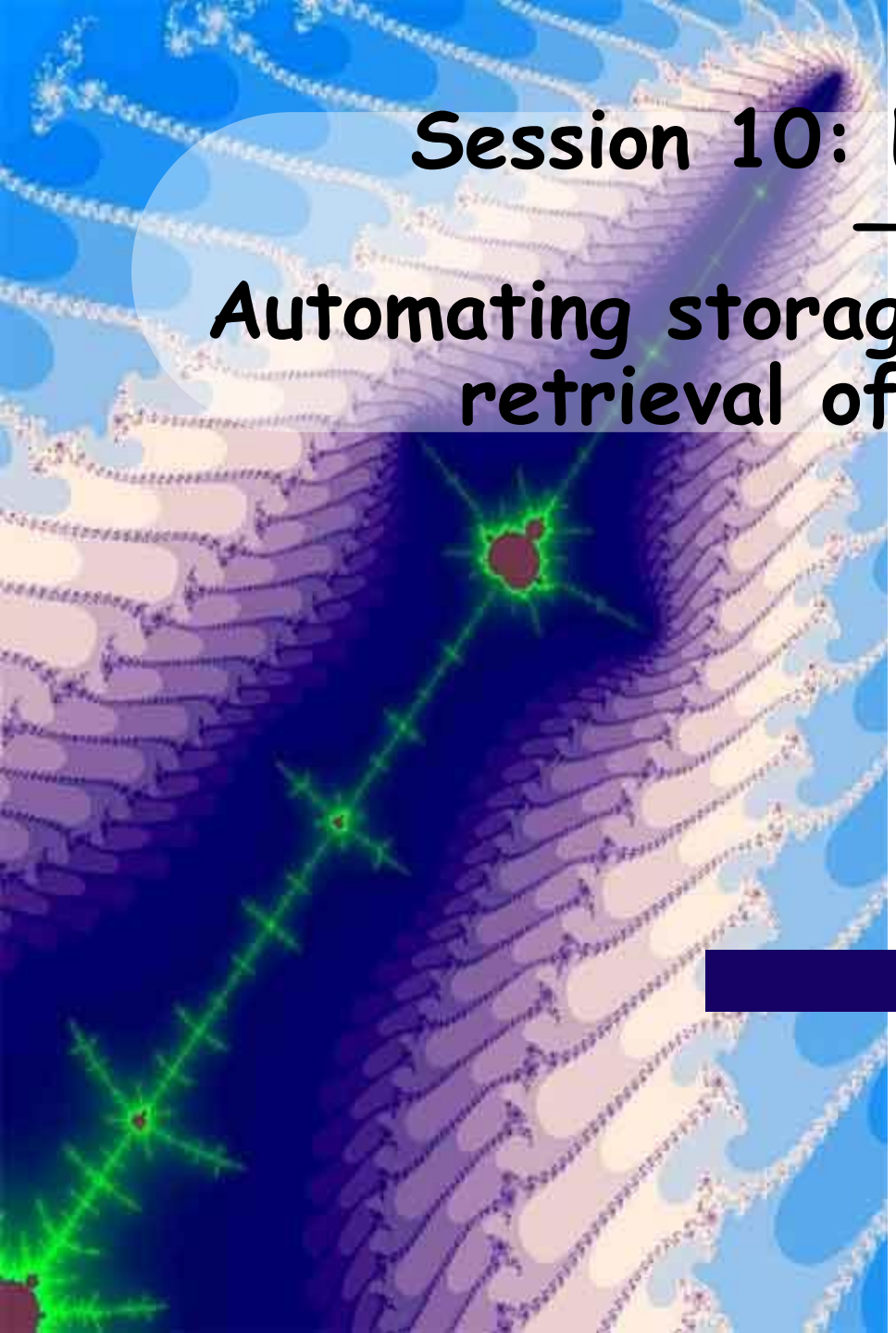


Session 10: Episode 3(2)

—

Automating storage, management & retrieval of knowledge

William P. Hall

President
Kororoit Institute Proponents and Supporters Assoc., Inc. - http://kororoit.org

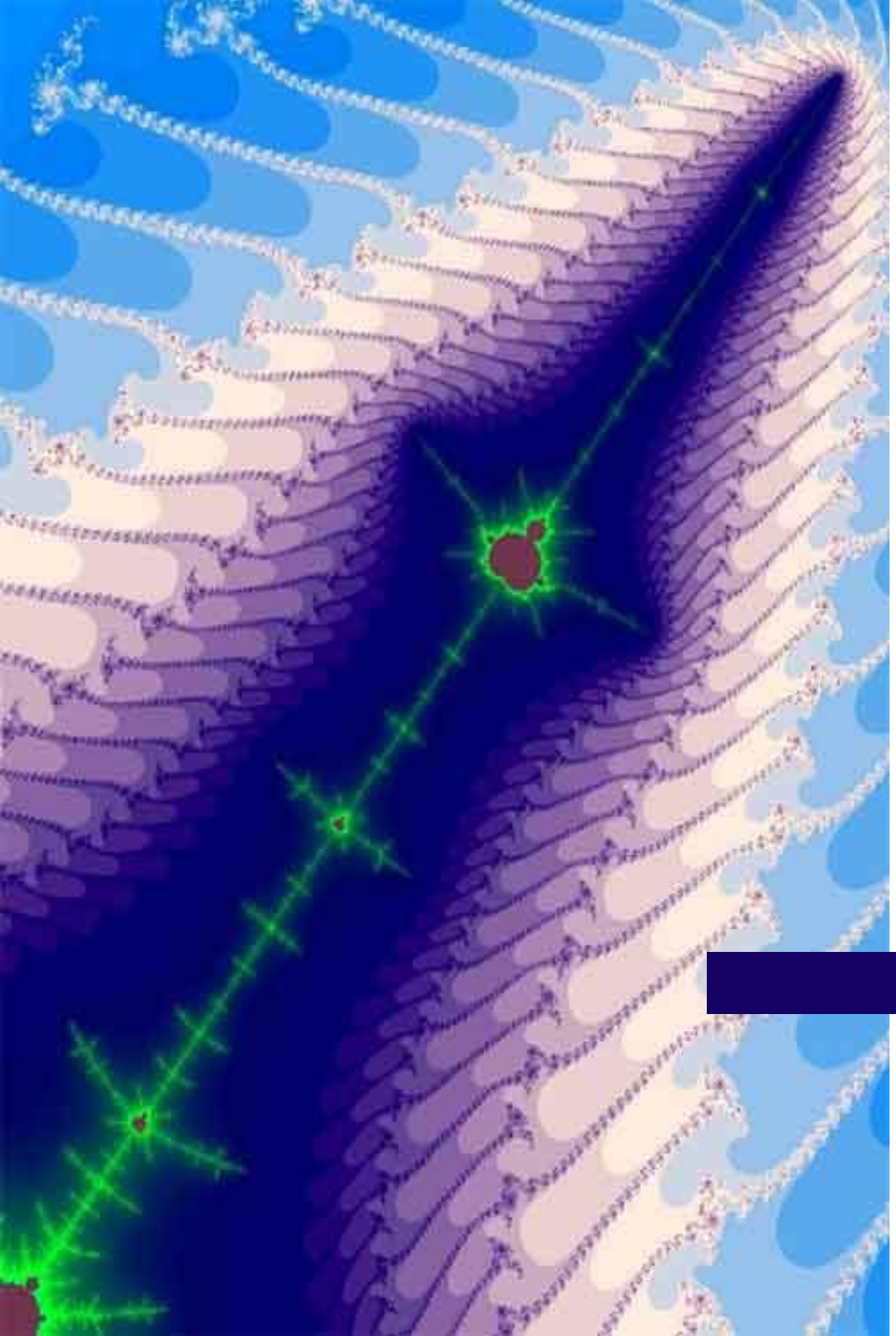
william-hall@bigpond.com
http://www.orgs-evolution-knowledge.net

Access my research papers from Google Citations

Tonight

- Last time we looked at how personal computing quantitatively extended human cognitive capabilities.
  - Most documents can now be filed, replicated and delivered at light speed to those who need the knowledge they contain.
  - These applications by themselves do not fundamentally change the cognitive activities of people assembling knowledge into documents or managing them
  - Most business processes still resemble those followed when scribes and clerks pressed cuneiform script onto clay tablets – they just work faster, more accurately, and with a lot less people
- Tonight we begin to explore how the external preservation and processing of knowledge extends cognition beyond single individuals to social and automated systems.

Episode 3(2) – Cognitive Tools for Individuals
Tools to Store, Manage and Retrieve Preserved Knowledge
Information science: disseminating, indexing and retrieving scholarly, scientific and technical knowledge
Computerizing and moving the indexes on-line
Indexing and semantic retrieval
The increasing cost of publishing paper and the limitations of libraries
The research library is dead – long live the world library



Information Science

Information science is concerned with analysing, collecting, classifying, manipulating, storing, retrieving, disseminating, and protecting information and knowledge.

Practitioners study the application and usage of knowledge in organizations, along with the interaction between people, organizations and any existing information systems, with the aim of creating, replacing, improving, or understanding information systems.

Information science is much more than computer systems and libraries

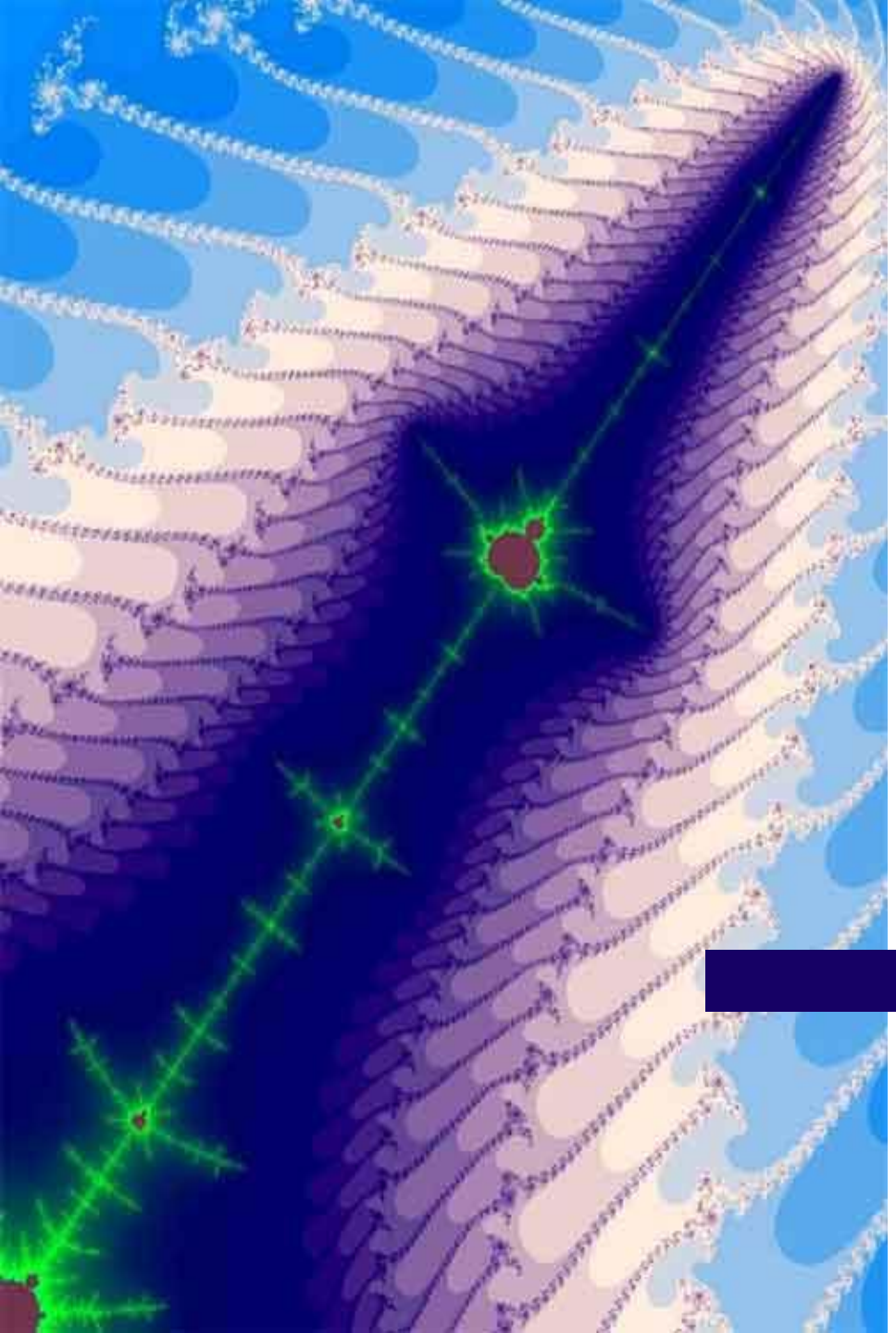
- ≠ computer science – IS's domain existed for thousands of years before computers were invented
- ≠ library science
  - IS's domain extends far beyond libraries, but
  - libraries are important information systems and today's main topic
- Information science is a handmaiden to scholarship and science
  - Its intellectual origins trace back to the ancient "libraries" and filing systems mentioned in Sessions 6 and 7
    - Royal Library of Ashurbanipal in the 7th Century BC in Assyria
    - Ancient Library of Alexandria that flourished from the 3rd to 1st Centuries BC
  - As soon as people began to accumulate stores of written knowledge for later use and sharing there was need to organize that storage so particular items/kinds of knowledge could be retrieved when needed or desired.

The Mouseion and Bibliotheka anticipate the modern university

- The Mouseion was a research and teaching institution highly reliant on the documents accumulated in the Bibliotheka
  - Aimed to have a (scribal) copy of every "book" in the world
  - Holdings estimated from a few tens of thousands of scrolls to more than a half million
  - Location and retrieval impossible without some form of logical filing system and indexing for retrieval
- Alexandrian cataloging system called the "Pinakes" provided
  - Subject index
  - Location
  - Bibliographic information about authors

Deficiencies of library catalogues for contents of complex publications were first recognized in Alexandria

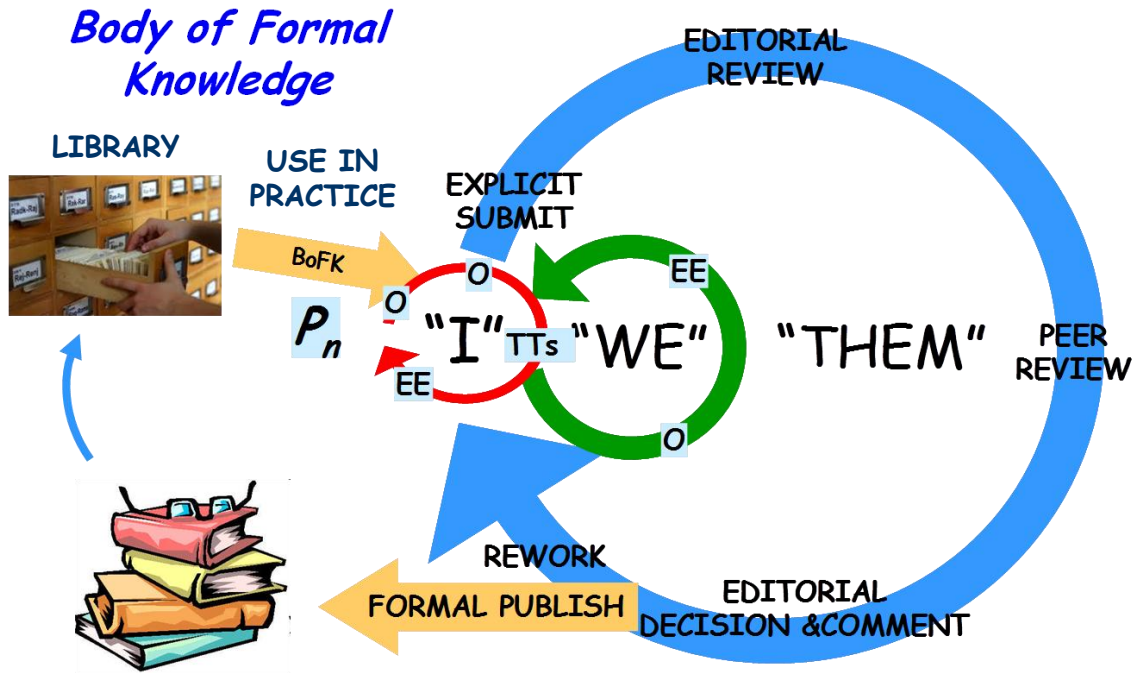
- Problems indexing volume containing many different articles by different authors
  - Callimachus (early librarian in the Bibliotheka) - '*Megabiblion, megakakon*' (big book, big evil)
  - Thomas Jefferson (whose library formed nucleus of US Library of Congress) - *for it is often doubtful to what particular subject a book should be ascribed. This is remarkably the case with books of travels, which often blend together the geography, natural history, civil history, agriculture, manufacturing, commerce, arts, occupations, manners, etc. of a country, so as to render it difficult to say to which they chiefly relate. Others again are polygraphical in their nature, as encyclopedias, magazines*
- The rise of "secondary literature" as "finding aids"
  - Reviews, etc. referencing papers relating to a particular subject
  - Subject indexes
  - Abstracting journals
  - Text books
  - Bibliographies



Continued growth and accumulation of scholarly & scientific journals made it ever more difficult for scholars and librarians to find particular information / knowledge

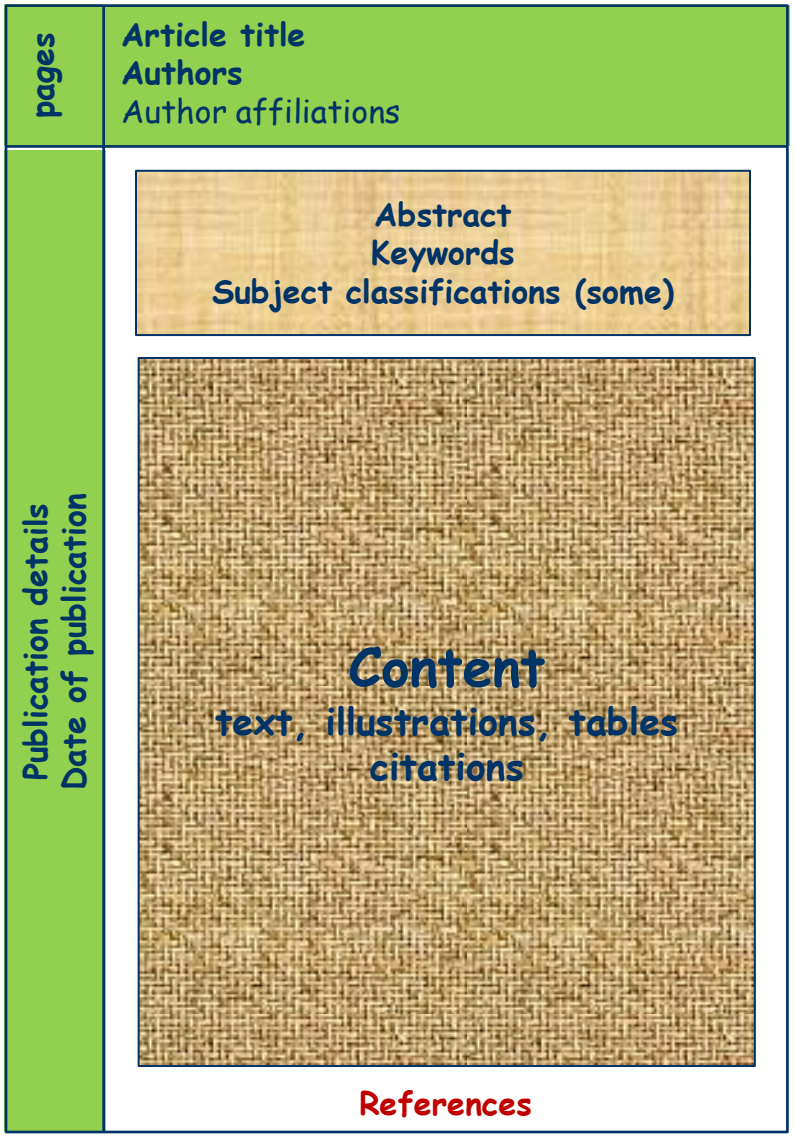
Cultural accumulation of explicit knowledge tended to be exponential – creating monstrous problems for libraries

- Science builds on & adds to prior knowledge
  - Deeper analysis
  - New theories
  - New disciplines
- For 200+ years libraries provided primary tools for finding/accessing prior knowledge



- Needed new ways to cope with increasing flood of journals and numbers of issues/articles per journal
  - Beyond the capacity of the library card catalog to index journal articles
  - Beyond the capacity of the individual researcher to scan all journals that might include relevant or important articles

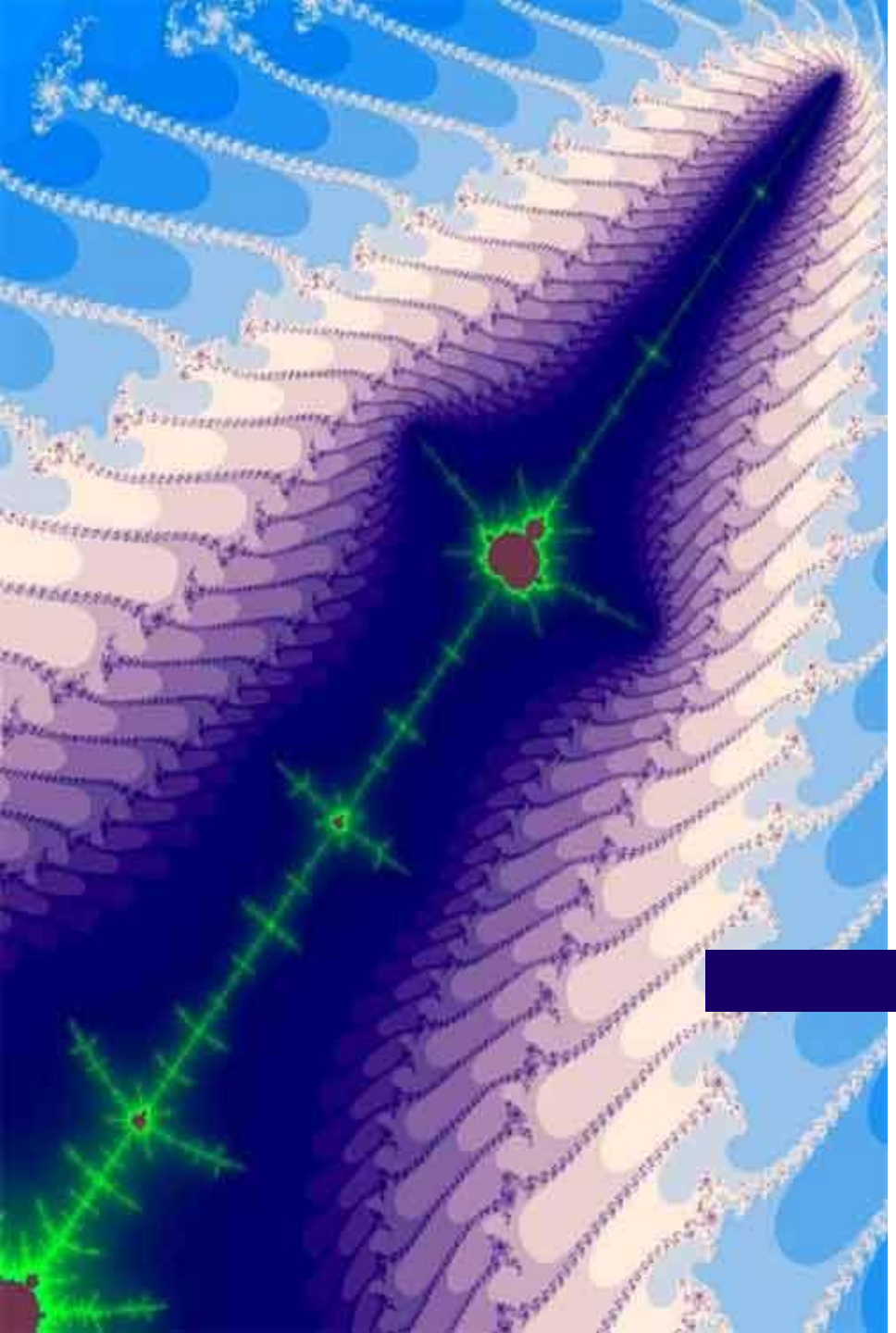
Scholarly/scientific articles are highly structured



- Basic framework of content
  - Statement of problem
  - Relations to other work
  - Description of original work
  - Discussion and interpretation
  - Conclusions
- Paragraphs
  - Narrative statements
  - Citations to references
    - prior work in discipline
    - methodologies & tools
    - sources of ideas
    - sources of additional evidence
- Supporting illustrations & tables

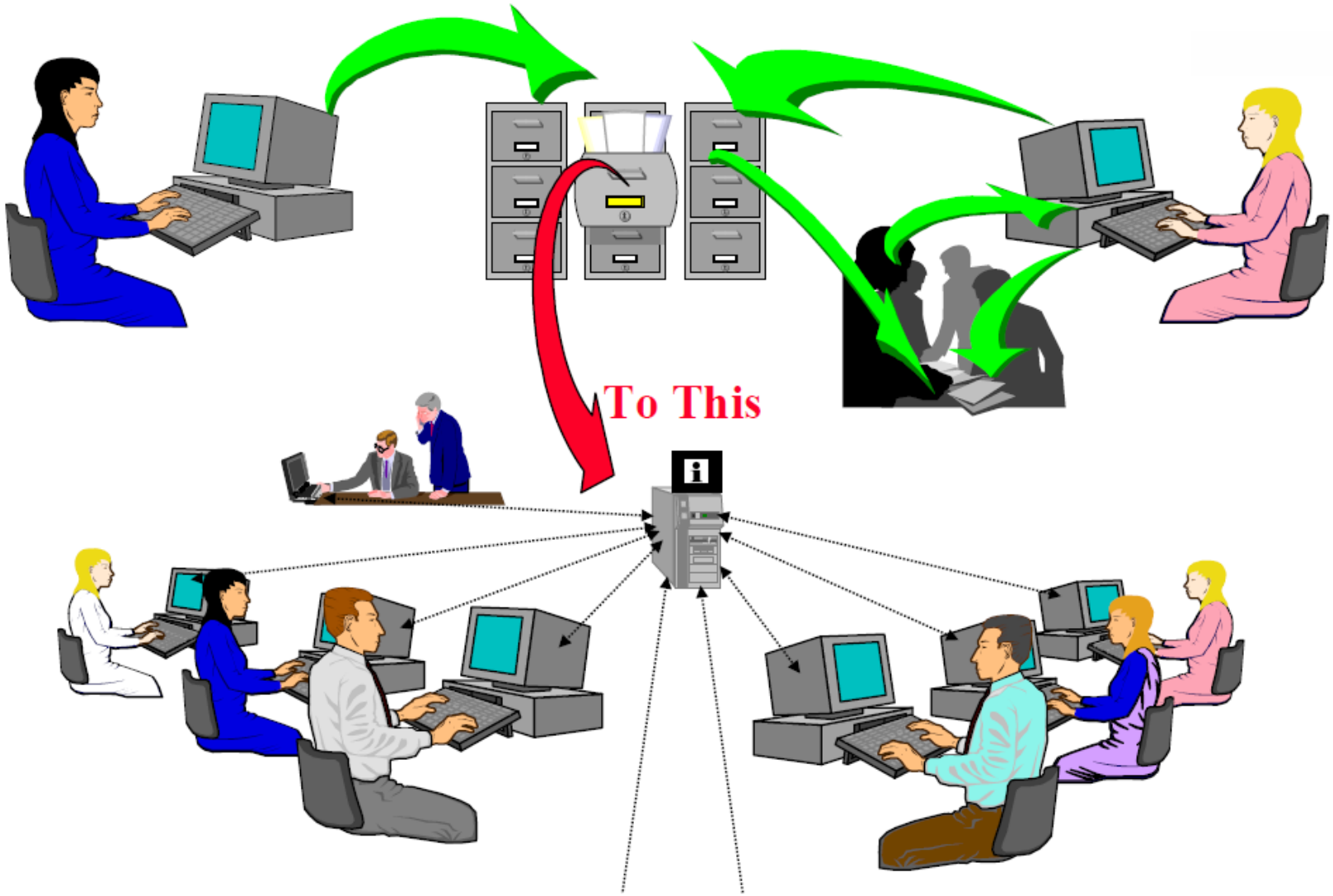
Emergence of paper-based commercially produced "finding aids"

- Reuss Repertorium, published from 1801–1821; Royal Society London Catalogue of Scientific Papers 1867-1925
  - Tried to be universal but overwhelmed by growth
- Disciplinary indexes with narrower foci
  - Individual indexers able to focus on subdisciplines they understand
  - Sciences
    - Zoological record 1865 until merged Biological Abstracts 1980
    - Index Medicus estab. by John Shaw Billings 1879 → MEDLINE/PubMed
    - Science (Physics) Abstracts 1889 →
    - Chemical Abstracts 1907 →
    - Biological Abstracts (BIOSIS) 1926 →
  - Legal
    - Shepards 1873
- Difficulties of manual production, printing & distribution
  - Hundreds of thousands of new papers to index every year
  - Required *many volunteers* to do abstracts to be remotely economical
  - 1-3 years behind the present before published



Beginnings of computerization: new and old indexes

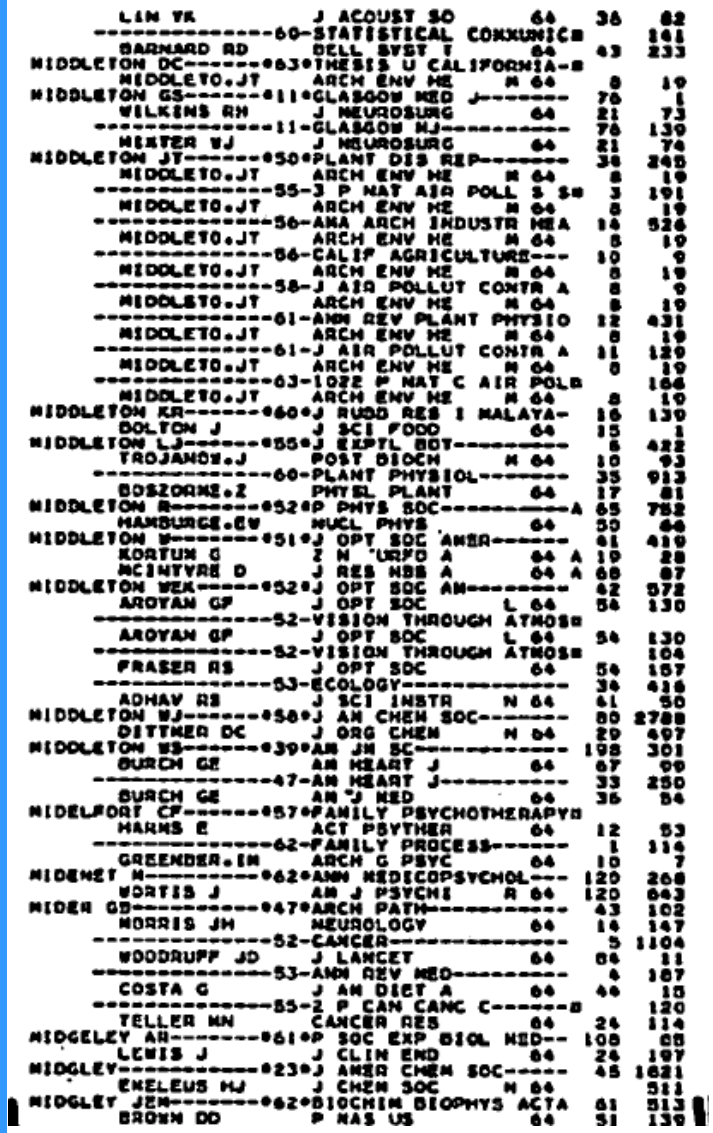
Computerization of indexes and documents changes the paradigm



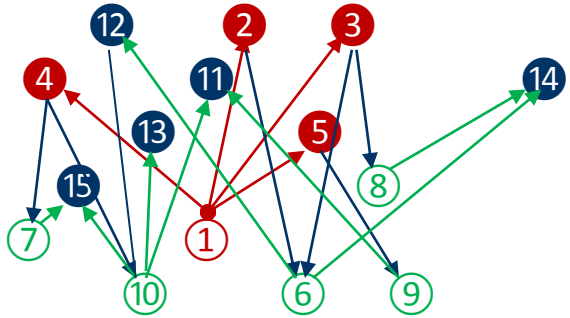
Computer generated indexes (~1960-80)

- Computer generated paper indexes from article metadata / abstracts
  - Card catalog approach for journal articles: author, title, subject point to complete citation (generally by ID#)
  - Keywords in context (words in titles, abstracts)
- Computer mediated finding aids enabled precision retrieval
  - Boolean logic (AND, OR, NOT) using multiple keywords with nested logic
  - Stemming (truncation by wildcard)
  - Proximity operators (e.g., NEAR, Sentence, Para, etc.)
  - Ranking and relevance
  - Keywords in context
  - Search within results sets
  - Full document retrieval (when available by email or snail-mail)
- **Citation indexing uses semantic relationships !!**

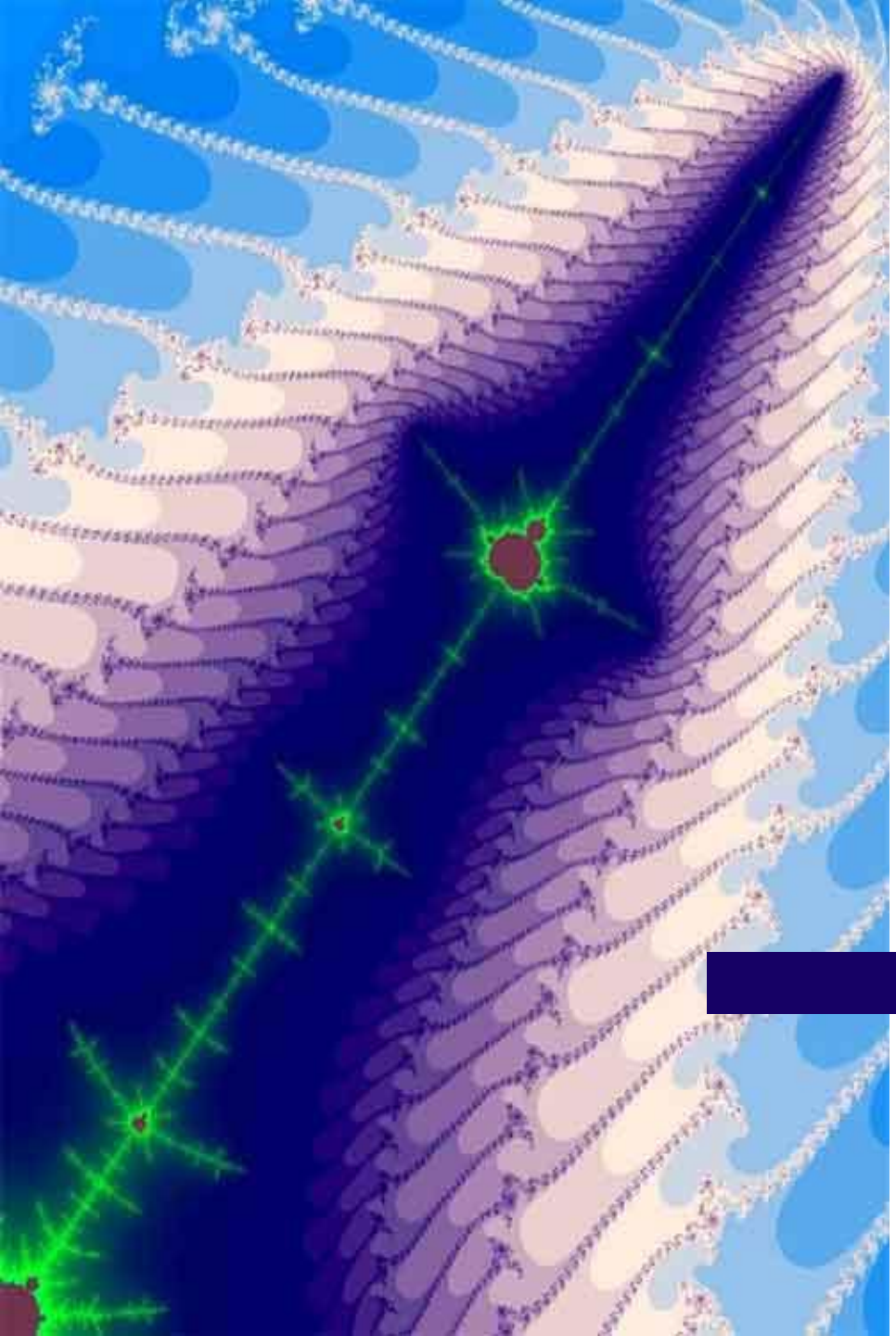
The semantic power of a new kind of indexing made feasible by computer processing



- Citation indexing
  - Use bibliographic metadata associated with every science article
  - Not readily managed by people
- Index of all articles citing a particular key reference
  - Semantics: Articles cite key reference because they use content from key reference
  - Articles citing key ref. probably contain more current, useful info



  - May go back and forth several times



Growth of scholarly & scientific journals created more difficulties for scholars and librarians

Statistics I collected for a 1999 ms: "Serving Scientific Knowledge to the Web"

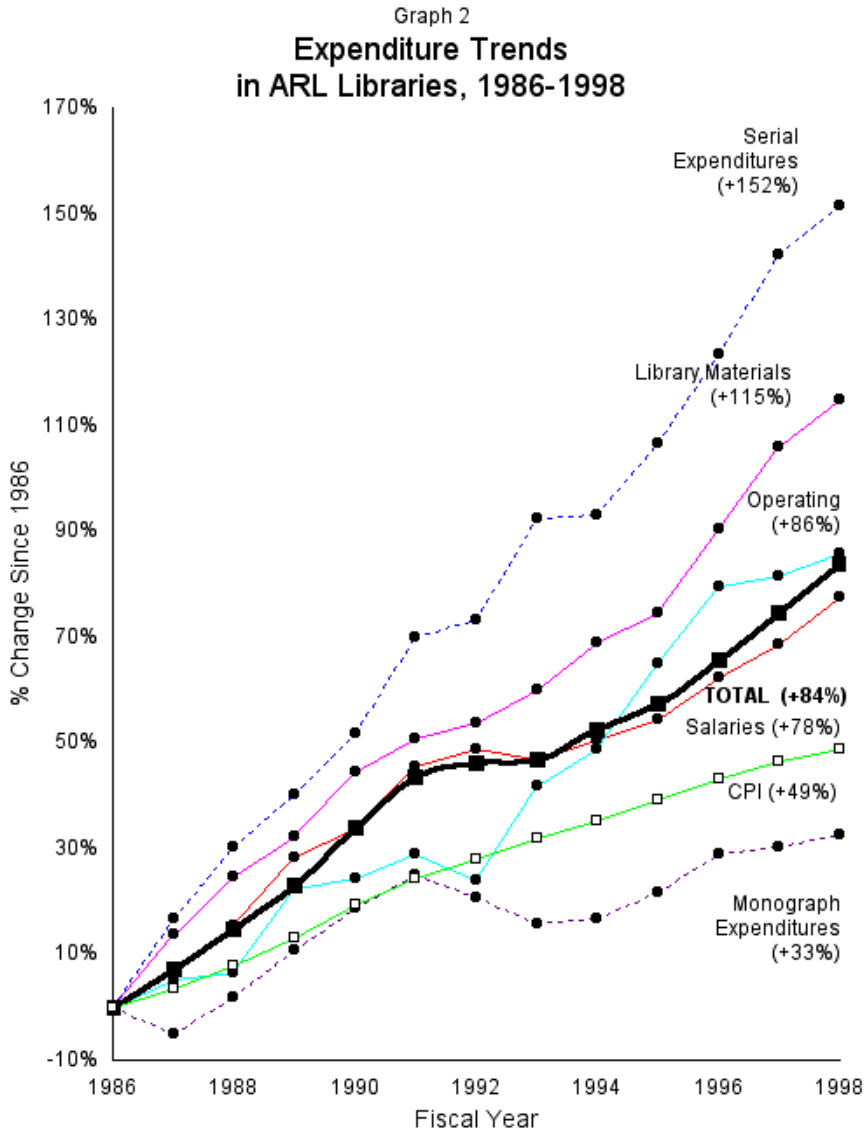
- BIOSIS
  - ~ 1997-1999 processed ~550,000 biomedical articles per year!
  - By 1997 it had accumulated > 13,000,000 article citations
- MEDLINE/PUBMED
  - Web interface accessing 9,000,000 journal citations in MEDLINE with links to full-text articles at participating publishers' Web sites
- Lexis-Nexis (legal & business)
  - 2.3 billion searchable documents
  - 9,862 databases
  - 6.8 million documents added each week
- How to find anything?

Growing volume of literature

- Increasing specialization in normal science
  - As volume of knowledge increased areas of competency became narrower and deeper
    - Can only find & read a limited number of pages a day
  - Increasing no. of disciplines & specialist journals
- Reduced turnaround time
  - Quicker access to prior knowledge
  - More rapid review & publishing enabled by email, word processing, & electronic publishing
- Insidious consequences of "publish or perish"
  - Academic survival depends on publishing more papers more quickly on more narrowly defined topics
  - More journals have to publish more papers every year

Statistics I collected for a 1999 ms: Serving Scientific Knowledge to the Web

- BIOSIS
  - ~ 1997-1999 processed ~550,000 biomedical articles per year!
  - By 1997 it had accumulated > 13,000,000 article citations
- MEDLINE/PUBMED
  - Web interface accessing 9,000,000 journal citations in MEDLINE with links to full-text articles at participating publishers' Web sites
- Lexis-Nexis (legal & business)
  - 2.3 billion searchable documents
  - 9,862 databases
  - 6.8 million documents added each week
- How to find anything?

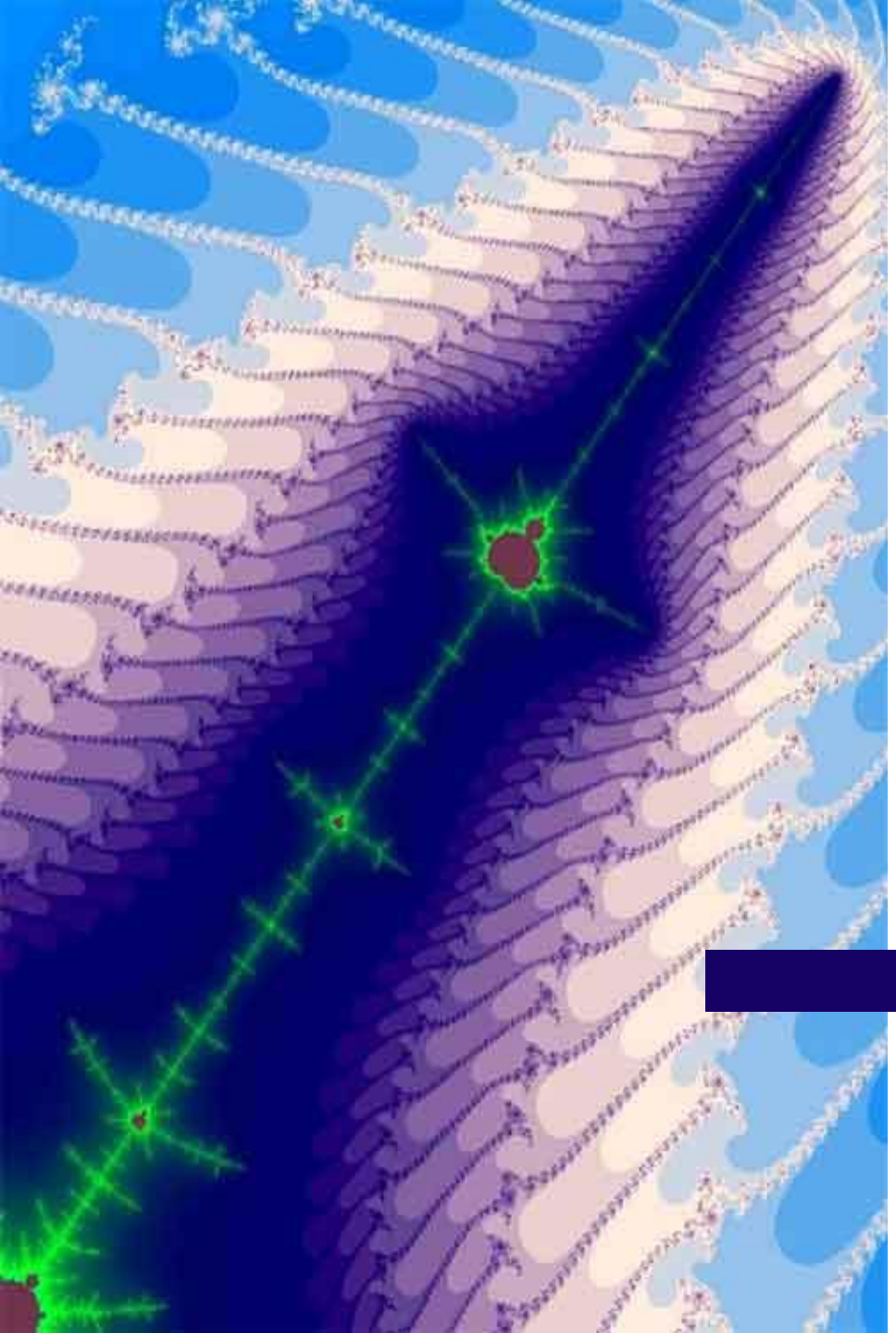


Price increases paid by two research libraries over three years (1995-1998)

SAMPLES OF SUBSCRIPTION PRICE INCREASES

	1995	1996	Change	1997	Change	1998	Change	Change 95 to 98
Brain Research	$10,181	$12,234	20.2%	$14,919	21.9%	$15,428	3.4%	51.5%
Biochim. Biophys. Acta	$7,555	$8,837	17.0%	$10,528	19.1%	$10,839	3.0%	43.5%
Chem. Phys. Letters	$5,279	$6,569	24.4%	$7,818	19.0%	$8,060	3.1%	52.7%
Eur. Jrnl. of Pharmacology	$4,576	$5,680	24.1%	$6,431	13.2%	$6,702	4.2%	46.5%
Gene	$3,924	$5,069	29.2%	$6,144	21.2%	$6,433	4.7%	63.9%
Inorganica Chim. Acta	$3,611	$4,476	24.0%	$5,283	18.0%	$5,540	4.9%	53.4%
Intl. Jrnl. of Pharmaceutics	$3,006	$3,915	30.2%	$4,691	19.8%	$4,983	6.2%	65.8%
Neuroscience	$3,487	$4,001	14.7%	$4,543	13.5%	$5,073	11.7%	45.5%
Theoretical Computer Science	$2,774	$3,425	23.5%	$3,835	12.0%	$4,059	5.8%	46.3%
Jrnl. of Exp. Marine Bio. & Eco.	$1,947	$2,445	25.6%	$2,811	15.0%	$2,931	4.3%	50.5%
Solid State Communications	$1,945	$2,327	19.6%	$2,602	11.8%	$2,871	10.3%	47.6%

- By 2000 the situation was so dire that University of California scholars threatened to boycott Nature Publishing Group over a proposed 400% increase in licensing fees



The WEB comes to the rescue

Moving the journals and indexes into online environments

- Much of the cost of publishing journals is in the cost of physically producing and delivering paper documents.
  - Remaining costs
    - Managing editor (may be volunteer)
    - Admin staff to manage the peer review process
    - Style editors
    - Electronic typesetters
    - Server and network administrators
  - Peers provide their reviews at no cost
  - Authors provide their copy in near-final electronic formats
- Libraries license and manage access to content on journal servers
  - Provide through connections to staff & subscribers to content managed on licensed journal servers
  - Many libraries no longer subscribe to paper journals
- Journal content becomes accessible to users as soon as loaded onto server
  - Reduces publication cycle from year or more to weeks
  - Great cost reduction possible

Accessing the Body of Formal Knowledge via the Web

- Web access to academic indexing services via library logins facilitates near instant access to relevant academic articles (gradual development since ~1990)
- Google Scholar trumps everything else
  - Released in beta in 2004
  - Khabsa & Giles 2014
    - > 114 M English language articles available on-line
    - Google Scholar indexes ~100 M
    - 24% available free to the Web
  - Enrique Orduña-Malea et al. 2014
    - Google Scholar indexes 160 M articles - > 3x Web of Knowledge
  - Provides wide variety of search tools
    - Key words, most Boolean, date limited, etc…
    - Citation indexing
  - Understands library portals and user's access via library
    - I can access everything Google knows about

The research library is dead – long live the world library ("global brain")

- The world library is the world body of formal knowledge